Econometrics Assignment 3

Ruoheng Du

05/15/2023

Since the beginning of China's reform and opening up, the country's economic development has generally been divided into four stages. The first two stages cover the periods of 1978-1992 and 1992-1997. During these stages, China made a historic transition from a low point in its economy to a market-oriented system, which posed tremendous challenges. The various economic entities within the country were not yet mature, and exploratory actions and pilot programs were necessary to determine the path of development. After 1998, things had changed. For example, China had entered into the World Trade Organization, which symbolized its integration into the wave of globalization, and indicated a significant shift in the country's economic landscape. Therefore, it is not feasible to generalize the first two stages of development with the later period. Therefore, I will use the data from 1998 till now.

Question 1

The data that I will use is International Trade: Exports: Value (Goods): Total for China (People's Republic Of). It is recorded as the growth rate from same period previous year.

```
df <- get_fred_series("XTEXVA01CNM659S", observation_start = "1998-01-01")

df <- df %>%
  rename('exp' = 'XTEXVA01CNM659S')
head(df,3)
```

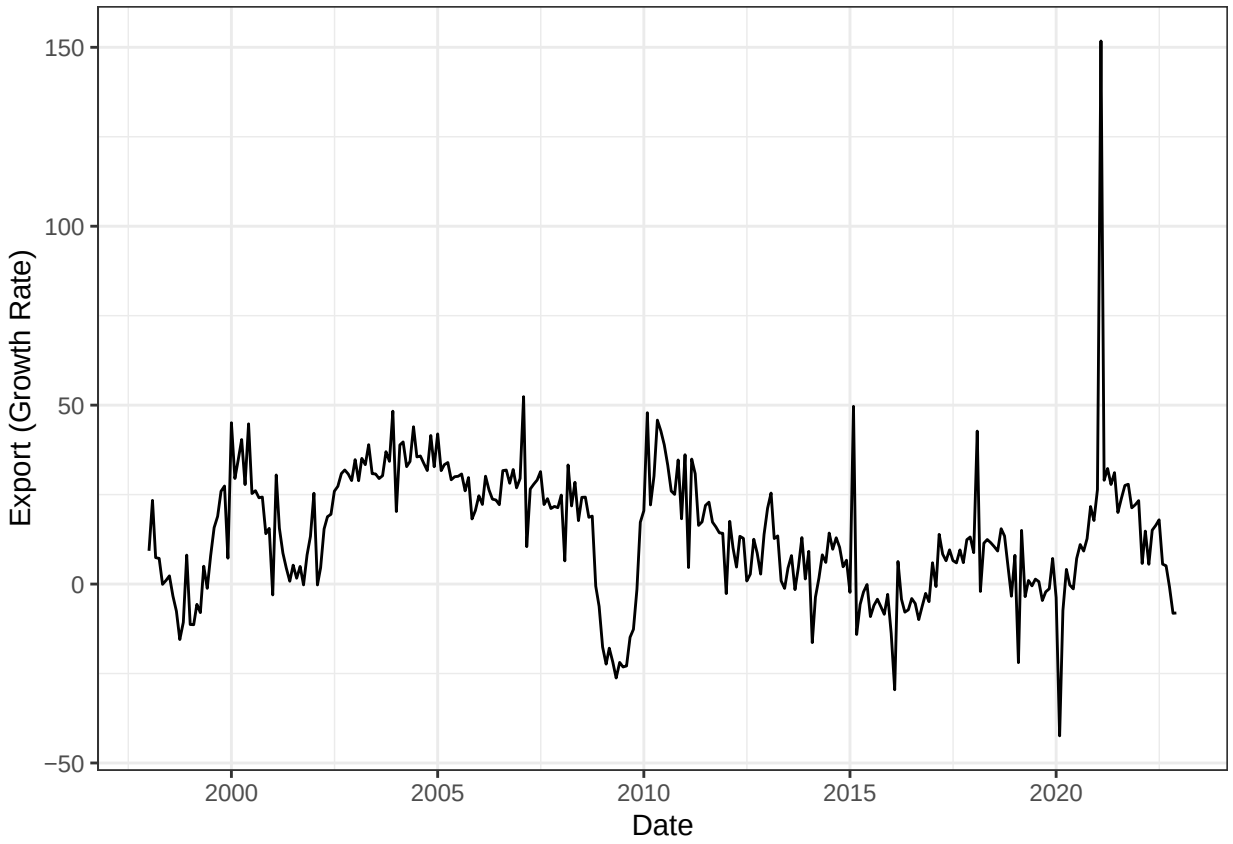
```
##         date       exp
## 1 1998-01-01  9.290350
## 2 1998-02-01 23.385823
## 3 1998-03-01  7.390392
```

```
tail(df,3)
```

```
##           date        exp
## 298 2022-10-01 -0.7619352
## 299 2022-11-01 -8.1625839
## 300 2022-12-01 -8.0836392
```

Question 2

```
ggplot(df, aes(date, exp))+
  geom_line() +
  theme_bw() +
  labs(x = 'Date', y = 'Export (Growth Rate)')
```

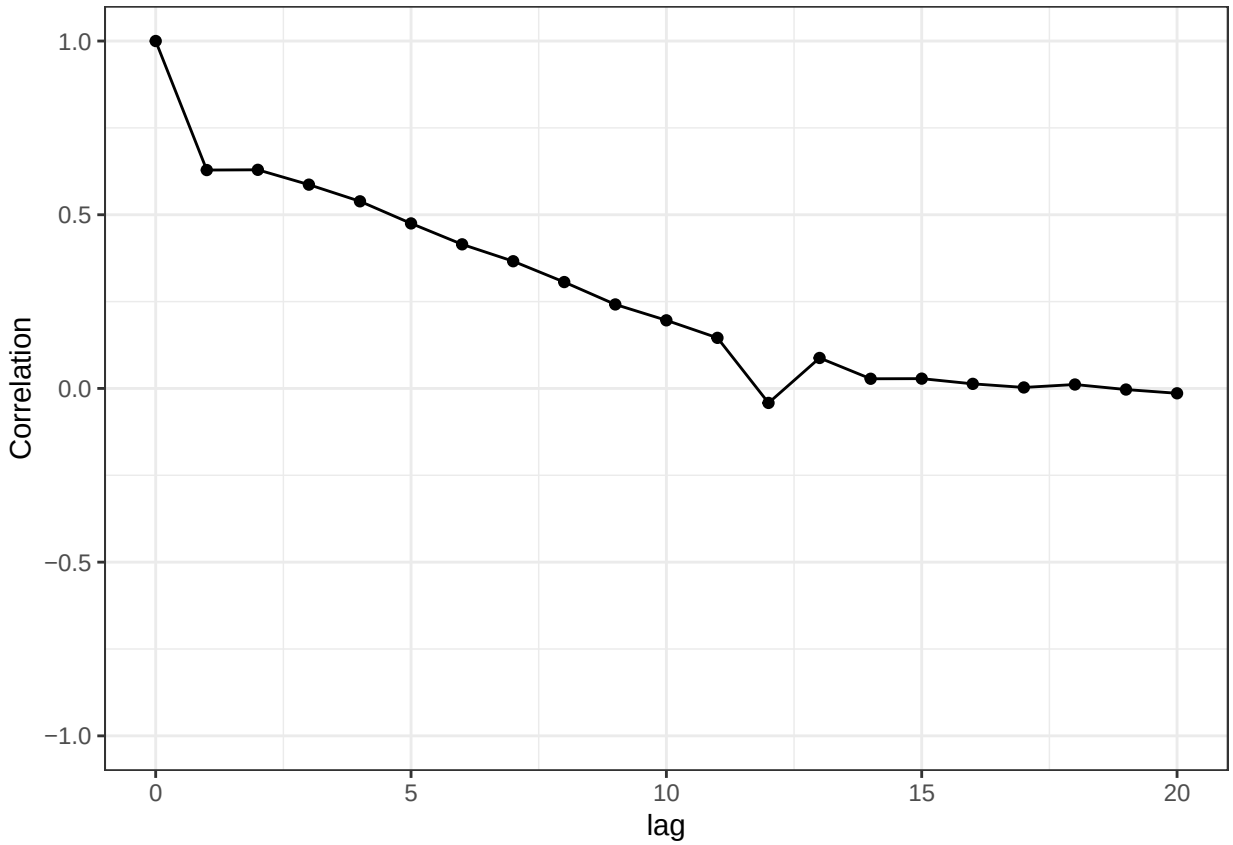


As we can see, starting from 1998, there are two periods of time that worth pointing out. • Around 2008: the financial crisis hit, causing great fluctuations in supply chain and creating strong negative outliers. • Around 2020: during the COVID-19, which is a huge supply shock, the export of international trade had faced strong negative shock at the beginning of the epidemic, while China's export industry greatly recovered after 1 year, resulting in a large positive outlier.

Question 3

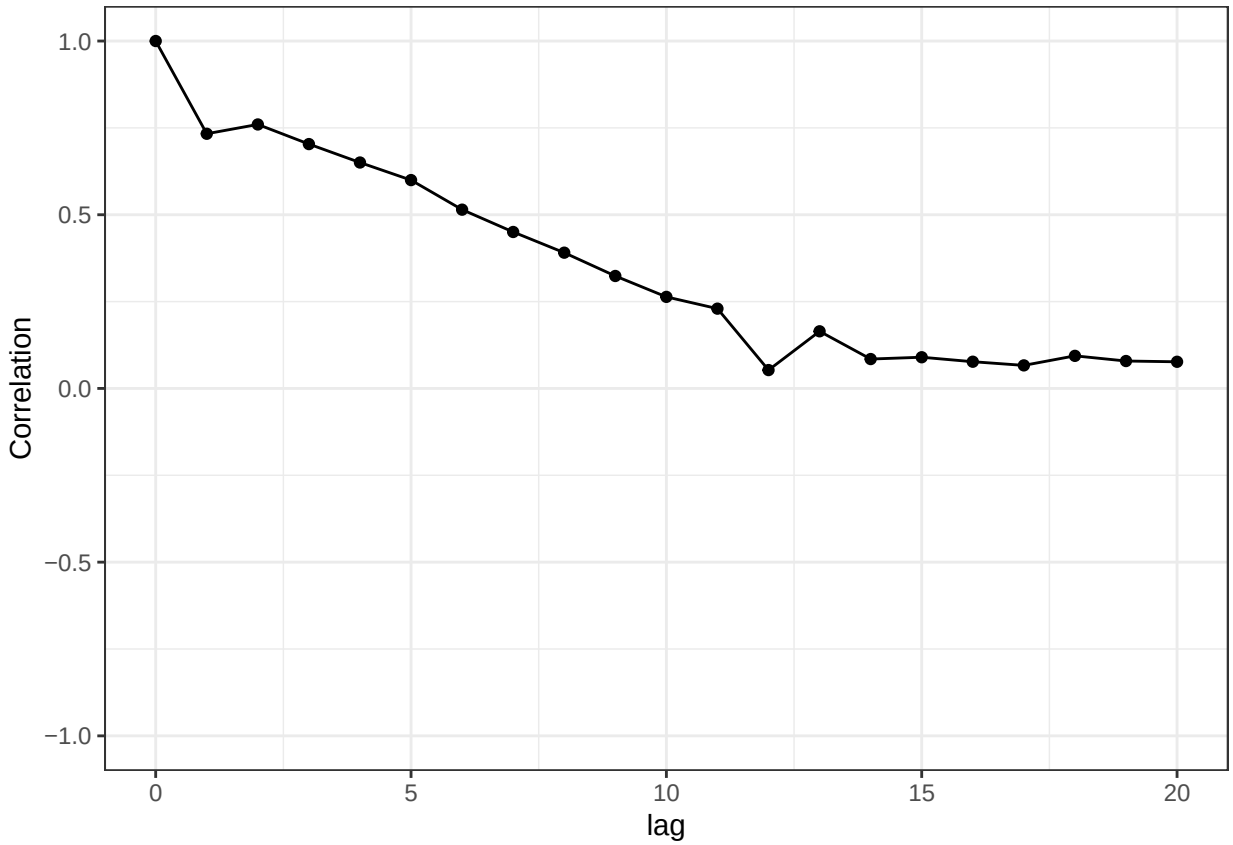
```
# Autocorrelation including COVID-19
x <- acf(df$exp, lag.max = 20, pl = FALSE)
acf_df <- data.frame(lag = 0:20, explag = x$acf)

ggplot(acf_df,aes(lag, explag)) +
  geom_point() +
  theme_bw() +
  ylim(-1,1) +
  geom_line() +
  labs(y = 'Correlation')
```




```
# Autocorrelation excluding COVID-19
df_no_covid <- df%>%filter(date < '2019-12-31')
x_no_covid <- acf(df_no_covid$exp, lag.max = 20, pl = FALSE)
acf_df_no_covid <- data.frame(lag = 0:20, explag = x_no_covid$acf)

ggplot(acf_df_no_covid,aes(lag, explag)) +
  geom_point() +
  theme_bw() +
  ylim(-1,1) +
  geom_line() +
  labs(y = 'Correlation')
```



The plot of ACF without the pandemic is almost identical to that with the pandemic, so, I will use the full dataset. There doesn't seem to have any seasonality and it seems that more recent time periods are better predictors because the correlation is higher. However, which AR is the most suitable is still unknown.

Question 4

```
#library(fUnitRoots)
#adfTest(df$exp)
adf.test(df$exp, nlag = 2)
```

```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##      lag   ADF p.value
## [1,]   0 -6.25    0.01
## [2,]   1 -3.70    0.01
## Type 2: with drift no trend
##      lag   ADF p.value
## [1,]   0 -8.16    0.01
## [2,]   1 -4.79    0.01
## Type 3: with drift and trend
##      lag   ADF p.value
## [1,]   0 -8.44    0.01
## [2,]   1 -4.99    0.01
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

From the ADF test, we can reject the null hypothesis that delta is equal to 0 at 1% significance level. As a result, the data is stationary and we should model the series in level.

Question 5

```
# AR1
fit1 = lm(exp ~ lag(exp), data = df)

#summary(fit1)
#BIC of AR1
BIC(fit1)
```

```
## [1] 2455.244
```

```
# AR4
fit4 = lm(exp ~ lag(exp) + lag(exp, 2) + lag(exp, 3) + lag(exp, 4), data = df)

summary(fit4)
```

```
##
## Call:
## lm(formula = exp ~ lag(exp) + lag(exp, 2) + lag(exp, 3) + lag(exp,
##     4), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.189  -5.454  -0.434   4.902 131.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19251    1.02382   2.142  0.03306 *
## lag(exp)     0.29357    0.05847   5.021 8.97e-07 ***
## lag(exp, 2)  0.29242    0.06009   4.866 1.87e-06 ***
## lag(exp, 3)  0.17567    0.05999   2.929  0.00368 **
## lag(exp, 4)  0.07682    0.05851   1.313  0.19022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.97 on 291 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.5142, Adjusted R-squared:  0.5075
## F-statistic:    77 on 4 and 291 DF,  p-value: < 2.2e-16
```

```
# BIC of AR4
BIC(fit4)
```

```
## [1] 2386.377
```

Since the BIC for the AR(4) model is smaller, we should use the AR(4) model. The goodness-of-fit for the AR(4) model is a R-squared value of 0.5142, adjusted R-squared value of 0.5075 and a BIC score of 2386.377.

Question 6

The data that I will use is Interest Rates: 3-Month or 90-Day Rates and Yields: Treasury Securities: Total for China (People's Republic Of), and it starts from 1998 (same for my previous dataset).

```
dfrate <- get_fred_series("IR3TTS01CNM156N", observation_start = "1998-01-01")
dfrate <- dfrate %>%
  rename('rate' = 'IR3TTS01CNM156N')
```

```
#library(fUnitRoots)
#adfTest(dfrate$rate)
adf.test(dfrate$rate, nlag = 2)
```

```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##      lag   ADF p.value
## [1,]   0 -2.24  0.0248
## [2,]   1 -2.22  0.0268
## Type 2: with drift no trend
##      lag   ADF p.value
## [1,]   0 -4.63    0.01
## [2,]   1 -4.38    0.01
## Type 3: with drift and trend
##      lag   ADF p.value
## [1,]   0 -4.68    0.01
## [2,]   1 -4.46    0.01
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

From the ADF test, we can reject the null hypothesis that delta is equal to 0 at 5% significance level. As a result, the data is stationary and we should model the series in level.

```
mdf = inner_join(df, dfrate, by = 'date')

adl1 <- lm(exp ~ lag(exp) + lag(exp, 2) + lag(exp, 3) + lag(exp, 4) +
             lag(rate), mdf)

#summary(adl1)
BIC(adl1)
```

```
## [1] 2382.053
```

```
adl4 <- lm(exp ~ lag(exp) + lag(exp, 2) + lag(exp, 3) + lag(exp, 4) +
             lag(rate) + lag(rate, 2) + lag(rate, 3) + lag(rate, 4), mdf)

summary(adl4)
```

```
##
## Call:
## lm(formula = exp ~ lag(exp) + lag(exp, 2) + lag(exp, 3) + lag(exp,
##     4) + lag(rate) + lag(rate, 2) + lag(rate, 3) + lag(rate,
##     4), data = mdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.826  -5.059  -0.302   4.941 132.050
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   7.05246    2.53150   2.786  0.00570 **
## lag(exp)      0.27324    0.05925   4.612 6.06e-06 ***
## lag(exp, 2)   0.27956    0.06049   4.621 5.80e-06 ***
## lag(exp, 3)   0.17236    0.06016   2.865  0.00449 **
## lag(exp, 4)   0.08158    0.05875   1.389  0.16606
## lag(rate)     0.28569    1.51038   0.189  0.85011
## lag(rate, 2)  0.82129    1.88060   0.437  0.66265
## lag(rate, 3) -0.83232    1.88206  -0.442  0.65865
## lag(rate, 4) -1.54481    1.48178  -1.043  0.29805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 283 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.5229, Adjusted R-squared:  0.5094
## F-statistic: 38.78 on 8 and 283 DF,  p-value: < 2.2e-16
```

```
BIC(adl4)
```

```
## [1] 2374.248
```

In this case, q = 4 is better since the BIC value of 2374.248 is smaller than that of 2382.053. This is better than the AR(4) model in the sense that this BIC value of 2374.248 is smaller than that of 2386.377, and this model has higher R-squared value and adjusted R-squared value. But we need to notice that the t-statistics of the last five coefficients are not statistically significant.

Question 7

```
forecast1 = as.numeric(coef(adl4)[1] + coef(adl4)[2] * tail(mdf$exp, n = 1) + coef(adl4)[3] * mdf[nrow(m

forecast1
```

```
## [1] 0.1775672
```

```
forecast2 = as.numeric(coef(fit4)[1] + coef(fit4)[2] * tail(mdf$exp, n = 1) + coef(fit4)[3] * mdf[nrow(m

forecast2
```

```
## [1] -2.308354
```

Question 8

```
predicted_value1 <- as.numeric(coef(adl4)[1] + coef(adl4)[2] * tail(mdf$exp, n = 1) + coef(adl4)[3] * mo

RMSFE1 <- summary(adl4)$sigma

lower_range1 <- forecast1 - 1.645*RMSFE1
upper_range1 <- forecast1 + 1.645*RMSFE1

bound_90_1 <- c(lower_range1, upper_range1)
cat('The 90% forecast interval for ADL(4,4) model forecast next period is:', bound_90_1)
```

```
## The 90% forecast interval for ADL(4,4) model forecast next period is: -21.20859 21.56373
```

```
predicted_value2 <- as.numeric(coef(fit4)[1] + coef(fit4)[2] * tail(mdf$exp, n = 1) + coef(fit4)[3] * mo

RMSFE2 <- summary(fit4)$sigma
```

```r
lower_range2 <- forecast2 - 1.645 * RMSFE2
upper_range2 <- forecast2 + 1.645 * RMSFE2

bound_90_2 <- c(lower_range2 ,upper_range2)

cat('The 90% forecast interval for AR(4) model forecast next period is:', bound_90_2)
```

```
## The 90% forecast interval for AR(4) model forecast next period is: -23.65064 19.03393
```